W774–W778 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki429

PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts

Thomas Goetz and Claus-Wilhelm von der Lieth*

Central Spectroscopy Department (B090)–Molecular, Modeling, German Cancer Research Center Heidelberg, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

Received February 14, 2005; Revised and Accepted March 24, 2005

ABSTRACT

Since it is becoming increasingly laborious to manually extract useful information embedded in the ever-growing volumes of literature, automated intelligent text analysis tools are becoming more and more essential to assist in this task. PubFinder (www.glycosciences.de/tools/PubFinder) is a publicly available web tool designed to improve the retrieval rate of scientific abstracts relevant for a specific scientific topic. Only the selection of a representative set of abstracts is required, which are central for a scientific topic. No special knowledge concerning the query-syntax is necessary. Based on the selected abstracts, a list of discriminating words is automatically calculated, which is subsequently used for scoring all defined PubMed abstracts for their probability of belonging to the defined scientific topic. This results in a hit-list of references in the descending order of their likelihood score. The algorithms and procedures implemented in PubFinder facilitate the perpetual task for every scientist of staying up-to-date with current publications dealing with a specific subject in biomedicine.

INTRODUCTION

To cope with an ever-increasing number of published scientific articles that are spread over a likewise growing number of scientific journals, new effective and easy-to-use tools are required to support scientists to stay up-to-date with the new findings in their specific area of research as well as to give a liable overview when entering a new scientific field (1–3).

The NCBI PubMed-Medline Service (www.ncbi.nlm.nih.gov/entrez/) provides a comprehensive resource of freely available scientific abstracts in the area of biomedical research. Therefore, it is well-suited to develop and test new approaches in text mining methodologies. Normal keyword searching is the most frequently applied way to query the Medline database (4). The retrieved titles and abstracts are most often manually screened. Useful abstracts are often buried in hundreds of less relevant papers and it is often not easy to evaluate if a retrieved abstract matches well with the desired topic. Even for very specific subjects, it is not always clear how to narrow the search to focus on the most relevant papers. Assuming a scientist is interested in the role of heparin as a potential drug in cancer treatment and queries the PubMed server with the terms 'heparin AND cancer', this search currently returns more than 2000 references, which encompass many biological events of heparin like inflammation, blood coagulation and cancer.

Several services exist where information provider like PubMed, BioMail, JADE, PubCrawler, OVID and ScienceDirect notify their users periodically of new publications, report literature or other sources of information on subjects in which the users have specified their interest. The performance of these services has been recently reviewed (5). XplorMed (6) (www.bork.embl-heidelberg.de/xplormed/) is a web tool that aids MEDLINE searching by summarizing the subjects contained in the results of a PubMed keyword search. Liu and Altman describe a program for incremental updates of a bibliography using the PubMed RELATED ARTICLES function. This approach is able to select a very narrowly defined set of articles from the literature (7).

The aim of the PubFinder service (www.glycosciences.de/tools/PubFinder) is to support scientists to retrieve a higher rate of relevant papers for the subject they are interested in and thus reduce the necessary time to sort out many less relevant articles. Basic requirements for the broad acceptance of such an approach are: (i) the search must be easy to set up, (ii) minimal user interaction should be required, (iii) no special knowledge concerning the query-syntax should be necessary and (iv) the whole procedure should be highly automized. To be able to optimize search queries, the service should provide tools to evaluate, why a certain abstract has been highly ranked. Additionally, the service should provide a convenient way to improve search queries based on previous runs.

*To whom correspondence should be addressed. Tel: +49 6221 424541; Email: vonderlieth@dkfz-heidelberg.de

METHODS

The statistical approach is used to scan abstracts taken from PubMed-Medline and to select and rank just those abstracts discussing a certain topic: first, a reference dictionary was generated containing the frequencies of the 100 000 most common words contained in all abstracts available through PubMed since 1990. In a second step, discriminating words are derived from a set of user defined abstracts discussing the topic of interest—the training set—that appear at unexpectedly high or low frequencies compared with the reference dictionary. The frequency of occurrence of each word contained in the training set is compared with the frequency in the reference directory. The first 100 words exhibiting the highest difference in occurrence between both data sets are assigned as discriminating words for a specific topic.

In a third step all selected abstracts are scanned for all discriminating words and scored for their likelihood of discussing the given topic. The applied algorithm is based on the approach described by Marcotte *et al*. (8) for detecting protein–protein interaction through literature mining. The user has the possibility to display the automatically derived list of discriminating words and their frequencies in the training set (see Table 1).

In a fourth step—after having scanned all defined volumes of abstracts—a hit list of references in descending order of their likelihood score is presented. A direct link to the corresponding PubMed page, information about authors, title and bibliographic data is given for each entry.

IMPLEMENTATION

The above outlined algorithm was written in Perl and is processed on a PC-Linux cluster that currently consists of 11 standard PCs, each equipped with two processors (1.67 MHz), resulting in 22 cluster nodes. Depending on the number of free processors a complete year of PubMed abstracts, which include roughly half a million abstracts, can be scanned in ~1–3 min. The procedure can be easily expanded for growing demand.

Table 1. The 20 words that most discriminate abstracts of the example training set ('literature mining', see Figure 1) from other non-related abstracts

Discriminating word	Word frequency in training set	Frequency in dictionary	ln *P*-Score
Abstracts	7.8e−03	7.7e−06	−331
Medline	6.6e−03	2.9e−05	−214
Articles	6.5e−03	4.1e−05	−192
Text	5.0e−03	1.8e−05	−167
Information	1.0e−02	4.0e−04	−163
Databases	4.6e−03	3.1e−05	−132
Database	5.1e−03	7.3e−05	−121
Mining	2.9e−03	5.9e−06	−109
Precision	3.9e−03	2.9e−05	−109
Recall	3.3e−03	2.2e−05	−97
Abstract	2.6e−03	6.1e−06	−96
Literature	5.5e−03	1.9e−04	−96
Names	2.4e−03	5.7e−06	−85
Extraction	3.7e−03	7.0e−05	−81
Biomedical	2.5e−03	1.1e−05	−79
Data	8.4e−03	1.3e−03	−65
Fields	2.8e−03	4.6e−05	−62
Automatically	1.9e−03	8.6e−06	−62
Title	1.7e−03	3.5e−06	−62
Mesh	2.1e−03	1.3e−05	−61

All administrative tasks to run the service are handled automatically or at least in a semi-automatic way, so that only little human supervision is required. To be able to efficiently scan all PubMed abstracts they were downloaded and stored in a local database (starting from 1990), skipping abstracts whose total length does not exceed 100 characters. New articles are received and inserted once per day by an automatic update script. This daily update process takes <5 min. The overall storage space required per stored PubMed year averages ~650 MB, including the abstracts' text, author information and additional data provided by PubMed.

The PubFinder web-service was started in fall 2004. A user registration is required to manage the access to queries and results at any time. Registration also guarantees that the access to personal data is protected. All scan results are stored and the short topic description is used to recall former queries at any time. An online help and a collection of FAQs (www.glycosciences.de/tools/PubFinder/help.php) provide a quick overview on how to use the PubFinder web service.

USER INTERACTION

The user has to provide a list of abstracts (indicated by PubMed-IDs, see Figure 1), which describe well the scientific topic to be looked for. We found, that at least 8–10 abstracts should be provided. However, the rule 'the more, the better' is of course true and the number of highly relevant articles will increase with increasing homogeneity of the provided set of abstracts. A very small number of selected abstracts or the choice of a thematically inconsistent set of abstracts normally results in low scores.

The most practical way to include PubMed-IDs is to scan for references using the text-mining tools provided by PubMed and copy-paste the IDs of references, which are regarded as relevant into the input sheet of the 'new-scan' window. After having selected a sufficient number of abstracts a meaningful title has to be provided, which will be subsequently used to identify and manage different queries.

Having submitted the search, the status of the scan process can be pursued by a progress meter, which indicates the percentage of already performed search time of the overall estimated computing time. Depending on the free resources on the PC-cluster and the number of volumes of PubMed abstracts to be scanned, a search may take several minutes up to 1 or 2 h. The current load of the PC-Server is always displayed. The user will be informed by email as soon as a search has been finished.

As result a hit list of references in descending order of their likelihood score is presented. The number of retrieved relevant papers depends on the intensity a certain scientific question is investigated and thus the number of available publications. If a search reveals only references having low scores, this finding is a clear indication that the selected set of publications is too inhomogeneous and does not mimic a scientific topic well. PubFinder offers the option to restart a search by indicating a certain number of references retrieved in a previous run. The rescan option has been shown as an efficient way to improve

Figure 1. Setting up a new scan with PubFinder's 'New Scan' dialogue. After having provided a short description and selecting the starting year of the query, the user has to enter a set of PubMed IDs, which represent the scientific topic to be scanned for. In this example we chose a set of 28 PubMed abstracts that deal with literature mining.

the relevance of the number of retrieved papers. A suitable procedure is as follows: in a first run 10–20 abstracts are selected, which are known to cover a certain scientific subject well. From the retrieved entries all newly found papers exhibiting high relevance are resubmitted for a second run. Our experience shows, that this procedure works very well as long as consistent sets of references are selected.

CONCLUSIONS

Confronted with an ever-increasing number of published scientific articles, new advanced, automatic and easy to use tools are required to improve the rate of retrieved publications that are relevant. The algorithms and procedures implemented in PubFinder are capable of facilitating the task of staying up-to-date with current scientific knowledge for specific subject in biomedicine. No special knowledge concerning the query-syntax is necessary. The required selection of a representative set of references that are central for a scientific topic should be easily manageable for scientists. Normally they are well aware of the key publications in their specific field and the required PubMed-Ids can be easily retrieved and copied to PubFinder. In case researchers want to gain an overview in a new field, they can use the information extraction procedures provided by PubMed and thus select a preliminary set of relevant papers. The PubFinder option to restart searches allows—beginning with a crude selection of papers—to indicate more relevant articles from the retrieved references and thus optimize the set of relevant discriminating words consecutively. A rescan needs only modest user interaction—just an indication of relevant papers—as the rest of the procedure runs completely in an automatic mode. Additionally, to support the optimization process, PubFinder displays a complete list of retrieved discriminating words and their ranking, which may help to evaluate how to optimize a query.

With the RELATED ARTICLES function PubMed offers a similar service, which also uses a purely statistical approach. It is able to select a very narrowly defined set of articles from the literature (9). The PubMed RELATED ARTICLES function calculates the similarity of a given article with all other documents in PubMed mainly evaluating on the occurrence of terms in two articles and the rate of all abstracts, which contain the certain term. To allow an interactive response time, closely related documents are pre-computed for each document in PubMed so that the service has only to recall this list. As described above, PubFinder uses discriminating words, derived from a user definable set of articles relevant for a specific topic, which are subsequently used to score all other abstracts.

As can be expected from the different kind of scoring functions applied, both approaches produce varying but complementary results. A detailed analysis and comparison of both approaches will be published elsewhere. In practice it has turned out to be a very useful procedure, to define a set of relevant articles to start a PubFinder run by using the RELATED ARTICLES function. The immediate response time of the NCBI service is definitively one of its strong points. However, this gain in speed has to be paid by an inflexibility,

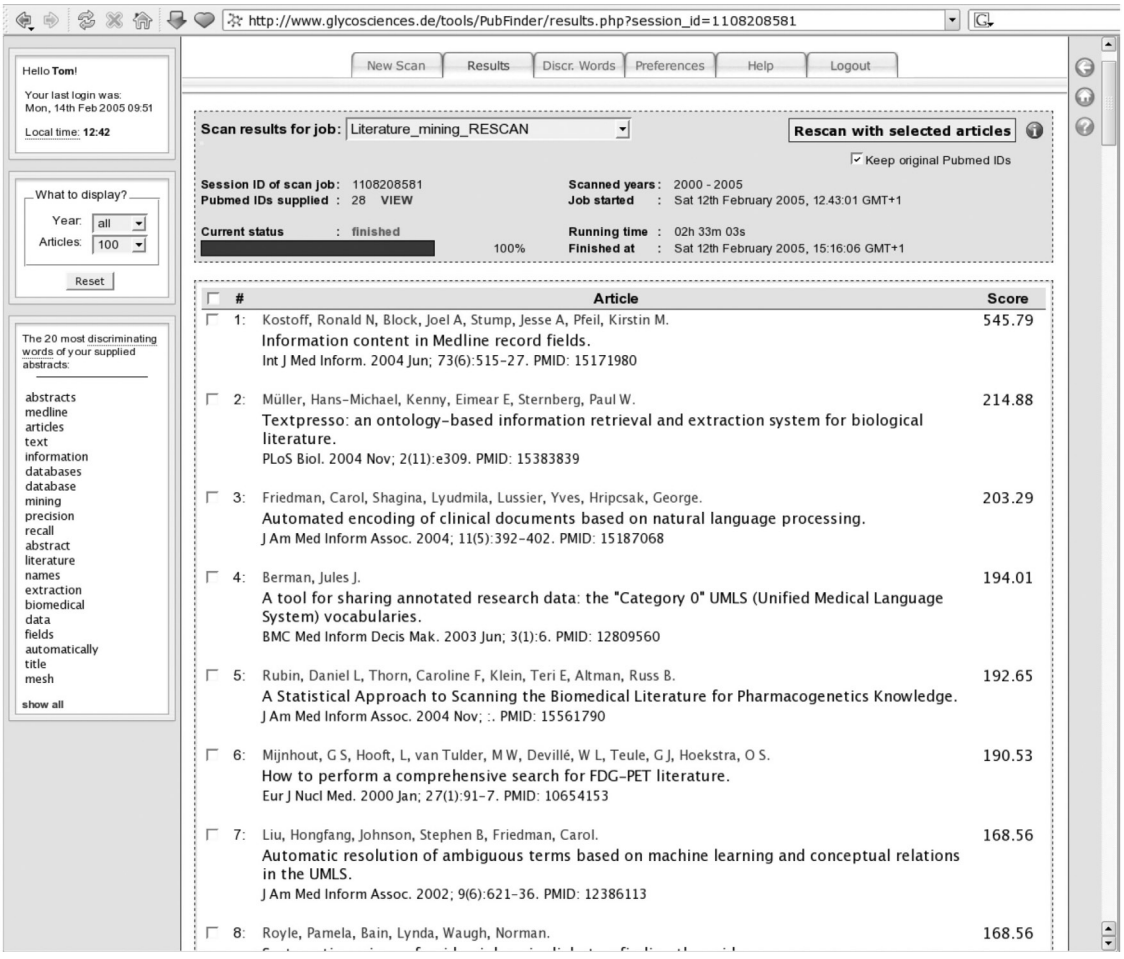

Figure 2. The scanned articles are listed in descending order according to their likelihood of discussing the user's given topic. The score is calculated as stated in (8) and represents the logarithm of the likelihood of belonging to the topic or not.

which does not allow influencing the criteria to retrieve similar articles. In contrast, the user interoperability of PubFinder is one of its strong points allowing users to subsequently optimize the search queries according to their specific needs. This method may also assist curators in acquiring pertinent literature in building biomedical databases dealing with specific topics (7).

The list of discriminating words generated by PubFinder may also help to support the process of deriving a catalogue of meaningful objects describing a domain of scientific interest as required to set an ontology (10).

A disadvantage of the PubFinder approach is its high demand of computing time. The most intensive part computationally is the scanning and scoring of abstracts. The required computer power increases linearly with the number of abstracts to be evaluated. Therefore, the initial derivation of an optimal set of discriminating words, which demands scanning several volumes of PubMed abstracts, may require a considerable amount of computing time. However, to keep up-to-date would not take up much time as only a relatively small amount of newly published abstracts would have to be reviewed.

ACKNOWLEDGEMENTS

The development of PubFinder is funded by a grant from the German Research Council (Deutsche Forschungsgemeinschaft: DFG) within the digital library program. Funding to pay the Open Access publication charges for this article was provided by the German Research Council (DFG).

Conflict of interest statement. None declared.

REFERENCES

1. Andrade,M.A. and Bork,P. (2000) Automated extraction of information in molecular biology. *FEBS Lett*., **476**, 12–17.
2. Chaussabel,D. (2004) Biomedical literature mining: challenges and solutions in the 'omics' era. *Am. J. Pharmacogenomics*, **4**, 383–393.
3. Rubin,D.L., Thorn,C.F., Klein,T.E. and Altman,R.B. (2005) A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge. *J. Am. Med. Inform. Assoc*., **12**, 121–129.
4. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Exploring MEDLINE abstracts with XplorMed. *Drugs Today (Barc)*, **38**, 381–389.
5. Shultz,M. and De Groote,S.L. (2003) MEDLINE SDI services: how do they compare? *J. Med. Libr. Assoc*., **91**, 460–467.

6. Perez-Iratxeta,C., Perez,A.J., Bork,P. and Andrade,M.A. (2003) Update on XplorMed: a web server for exploring scientific literature. *Nucleic Acids Res.*, **31**, 3866–3868.
7. Liu,X. and Altman,R.B. (1998) Updating a bibliography using the related articles function within PubMed. *Proc. AMIA Symp.*, 750–754.
8. Marcotte,E.M., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.
9. Perez-Iratxeta,C., Astola,N., Ciccarelli,F.D., Sha,P.K., Bork,P. and Andrade,M.A. (2003) A protocol for the update of references to scientific literature in biological databases. *Appl. Bioinformatics*, **2**, 189–191.
10. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: anontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.